TribeFlow: Mining & Predicting User Trajectories

Flavio Figueiredo[1,2], Bruno Ribeiro[4,5], Jussara Almeida[3], Christos Faloutsos[5]
[1]UFCG - Brazil, [2]IBM Research - Brazil, [3]UFMG - Brazil, [4]Purdue University, [5]Carnegie Mellon University
{flaviov,jussara}@dcc.ufmg.br, ribeiro@cs.purdue.edu, christos@cs.cmu.edu

ABSTRACT

Which song will Smith listen to next? Which restaurant will Alice go to tomorrow? Which product will John click next? These applications have in common the prediction of user trajectories that are in a constant state of flux over a hidden network (e.g. website links, geographic location). Moreover, what users are doing now may be unrelated to what they will be doing in an hour from now. Mindful of these challenges we propose TribeFlow, a method designed to cope with the complex challenges of learning personalized predictive models of non-stationary, transient, and time-heterogeneous user trajectories. TribeFlow is a general method that can perform next product recommendation, next song recommendation, next location prediction, and general arbitrary-length user trajectory prediction without domain-specific knowledge. TribeFlow is more accurate and up to $413\times$ faster than top competitors.

Keywords

User Trajectory Recommendation; Latent Environments;

1. INTRODUCTION

Web users are in a constant state of flux in their interactions with products, places, and services. User preferences and the environment that they navigate determine the sequence of items that users visit (links they click, songs they listen, businesses they visit). In this work we refer to the sequence of items visited by a user as the user's trajectory. Both the environment and user preferences affect such trajectories. The underlying navigation environment may change or vary over time: a website updates its design, a suburban user spends a weekend in the city. Similarly, user preferences may also vary or change over time: a user has different music preferences at work and at home, a user prefers ethnic food on weekdays but will hit all pizza places while in Chicago for the weekend.

The above facts result in user trajectories that over multiple time scales can be non-stationary (depend on wall clock times), transient (some visits are never repeated), and time-heterogeneous (user behavior changes over time); please refer to Section 5 for examples. Unfortunately, mining non-stationary, transient, and time-heterogeneous stochastic processes is a challenging task. It would be easier if trajectories were stationary (behavior is independent of wall clock times), ergodic (visits are infinitely repeated), and time-homogeneous (behavior does not change over time).

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883059.

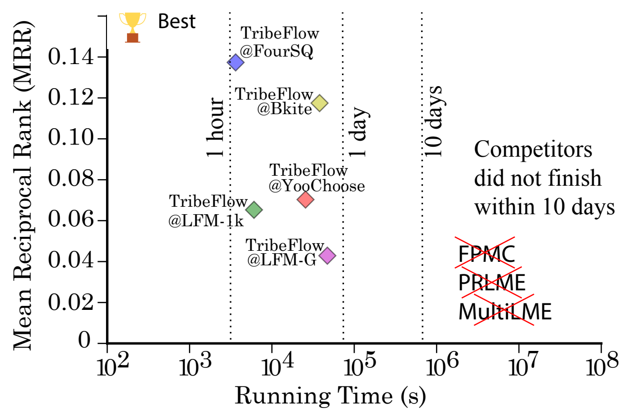

Figure 1: TribeFlow is at least an order of magnitude faster than state-of-the-art methods for next-item predictions.

In this work we propose TribeFlow to tackle the problem of mining and predicting user trajectories. TribeFlow takes as input a set of users and a sequence items they visit (user trajectories), including the timestamps of these visits if available, and outputs a model for personalized next-item prediction (or next $n > 1$ items). TribeFlow can be readily applied to personalized trajectories from next check-in recommendations, to next song recommendations, to product recommendations. TribeFlow is highly parallel and nearly two orders of magnitude faster than the top state-of-the-art competitors. In order to be application-agnostic we ignore application-specific user and item features, including time-of-day effects, but these can be trivially incorporated into TribeFlow.

To illustrate the performance of TribeFlow consider Figure 1, where we seek to compare the Mean Reciprocal Rank (MRR) of TribeFlow over datasets with up to 1.6 million items and 86 million item visits (further details about this dataset is given in Section 4) against that of state-of-the-art methods such as Multi-core Latent Markov Embedding (MultiLME) [40], personalized ranking LME (PRLME) [13], and Context-aware Ranking with Factorizing Personalized Markov Chains [45] (FPMC). Unfortunately, MultiLME, PRLME, and FPMC cannot finish any of these tasks in less than 10 days while for TribeFlow it takes between one and thirteen hours. In significantly sub-sampled versions of the same datasets we find that TribeFlow is at least 23% more accurate than its competitors.

TribeFlow works by decomposing potentially non-stationary, transient, time-heterogeneous user trajectories into *very* short sequences of random walks on latent environments that are stationary, ergodic, and time-homogeneous. An intuitive way to understand TribeFlow